

Introducción al Análisis de Regresión

Pasquini, Ricardo

IAE Universidad Austral

August 6, 2025

¿Para qué sirven los modelos de regresión?

- ▶ Predicción
- ▶ Extrapolación
- ▶ Inferencia causal
 - ▶ Experimentales
 - ▶ Observacionales

Predictión

Mapas de valor de la Tierra Urbana en Córdoba

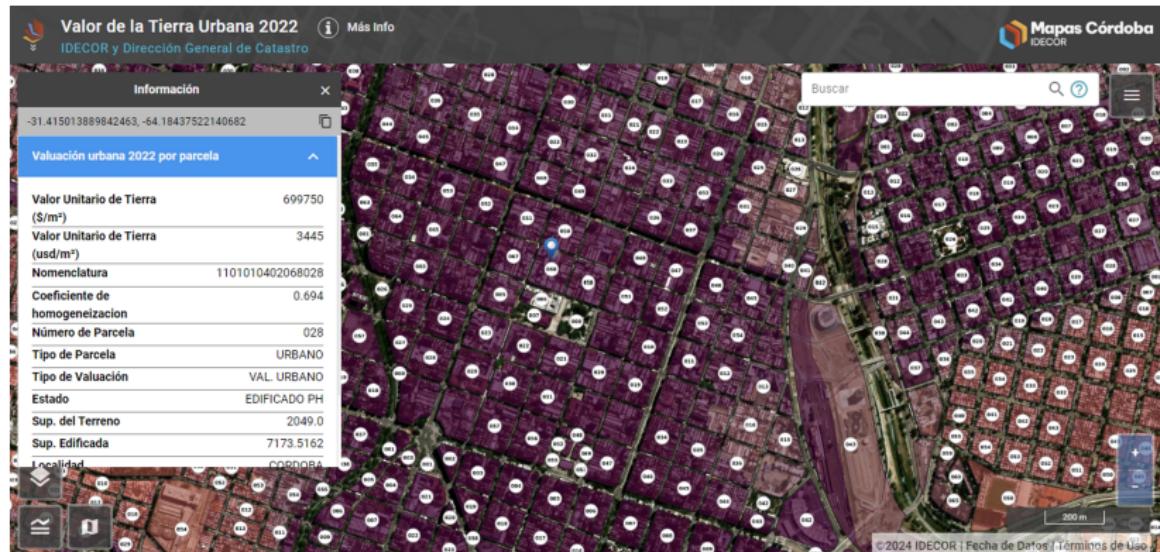


Figure: IDEDECOR <https://mapascordoba.gob.ar/viewer/mapa/401>

Predicción

Sistemas de recomendación

Startups

Airbnb Adds A Pricing Recommendation Tool For Renters

Matthew Lynley / 10:34 AM PDT • June 4, 2015

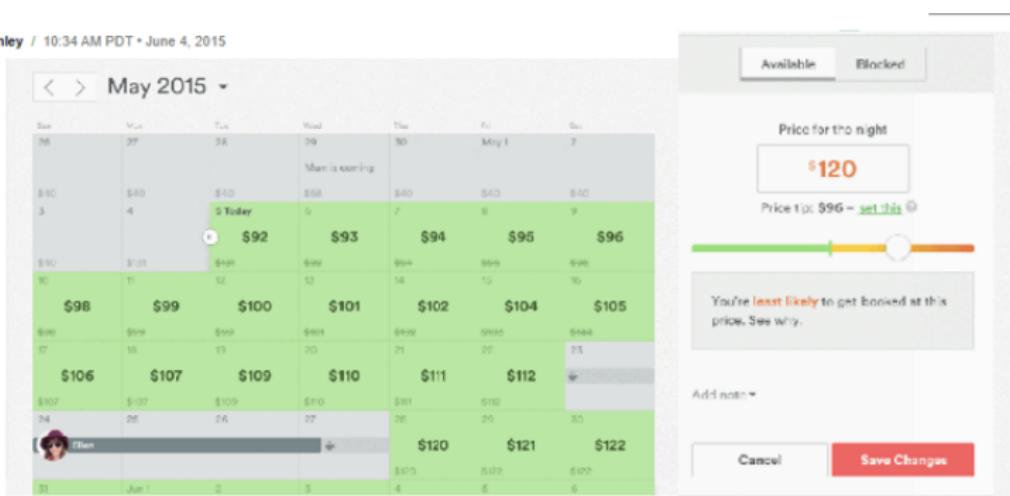


Image Credits: Airbnb

Figure: <https://techcrunch.com/2015/06/04/airbnb-adds-a-pricing-recommendation-tool-for-renters/>

Extrapolación

Regresión y Post-estratificación para predecir votación

Forecasting elections with non-representative polls

Wei Wang^{a,*}, David Rothschild^b, Sharad Goel^b, Andrew Gelman^{a,c}

ABSTRACT

Election forecasts have traditionally been based on representative polls, in which randomly sampled individuals are asked who they intend to vote for. While representative polling has historically proven to be quite effective, it comes at considerable costs of time and money.

Moreover, as response rates have declined over the past several decades, the statistical benefits of representative sampling have diminished. In this paper, we show that, with proper statistical adjustment, non-representative polls can be used to generate accurate election forecasts, and that this can often be achieved faster and at a lesser expense than traditional survey methods. We demonstrate this approach by creating forecasts from a novel and highly non-representative survey dataset: a series of daily voter intention polls for the 2012 presidential election conducted on the Xbox gaming platform. After adjusting the Xbox responses via multilevel regression and poststratification, we obtain estimates which are in line with the forecasts from leading poll analysts, which were based on aggregating hundreds of traditional polls conducted during the election cycle. We conclude by arguing that non-representative polling shows promise not only for election forecasting, but also for measuring public opinion on a broad range of social, economic and cultural issues.

Figure: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/forecasting-with-nonrepresentative-polls.pdf>

Extrapolación

Regresión y Post-estratificación para predecir votación

Forecasting elections with non-representative polls

Wei Wang^{a,*}, David Rothschild^b, Sharad Goel^b, Andrew Gelman^{a,c}

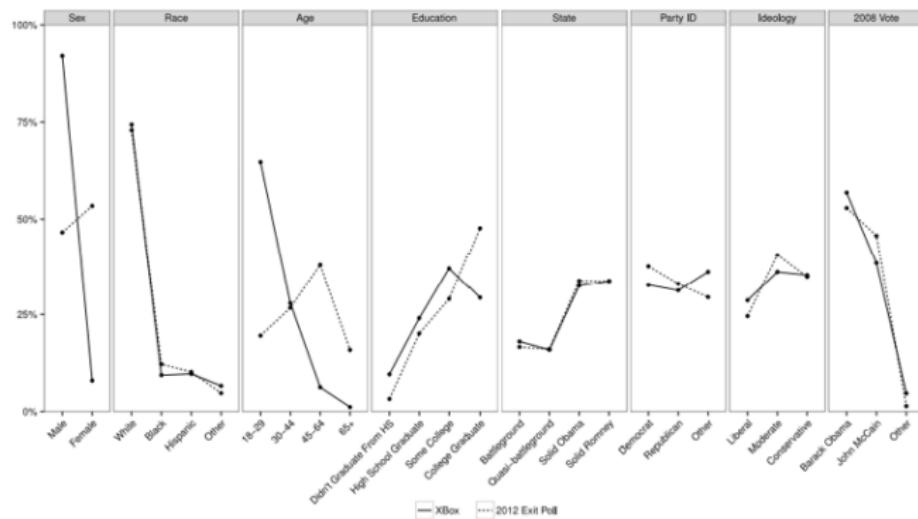


Fig. 1. A comparison of the demographic, partisan, and 2008 vote distributions in the Xbox dataset and the 2012 electorate (as measured by adjusted exit polls). As one might expect, the sex and age distributions exhibit considerable differences.

Figure: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/forecasting-with-nonrepresentative-polls.pdf>

Extrapolación

Regresión y Post-estratificación para predecir votación

Fig. 5. Comparison of the two-party Obama vote share for various demographic subgroups, as estimated from the 2012 national exit poll and from the Xbox data on the day before the election.

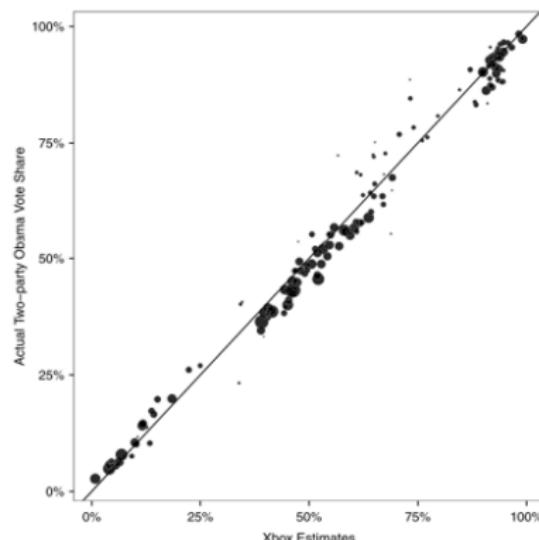
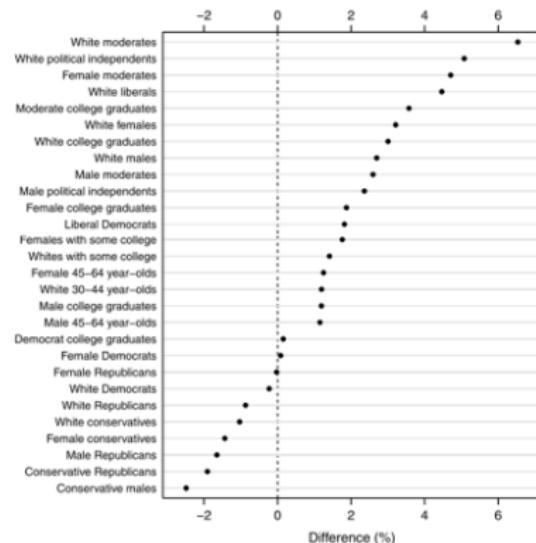


Figure: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/forecasting-with-nonrepresentative-polls.pdf>

Inferencia Causal - Experimental

Prácticas ESG y atracción de empleo

Polarizing Corporations: Does Talent Flow to "Good" Firms?

Emanuele Colonnelli, Timothy McQuade, Gabriel Ramos, Thomas Rauter, and Olivia Xiong

NBER Working Paper No. 31913

November 2023

JEL No. D2,G0,G3,G4,J0,O10,P0

ABSTRACT

We conduct a field experiment in partnership with the largest job platform in Brazil to study how environmental, social, and governance (ESG) practices of firms affect talent allocation. We find both an average job-seeker's preference for ESG and a large degree of heterogeneity across socioeconomic groups, with the strongest preference displayed by highly educated, white, and politically liberal individuals. We combine our experimental estimates with administrative matched employer-employee microdata and estimate an equilibrium model of the labor market. Counterfactual analyses suggest ESG practices increase total economic output and worker welfare, while increasing the wage gap between skilled and unskilled workers.

Figure:

https://www.nber.org/system/files/working_papers/w31913/w31913.pdf

Inferencia Causal - Experimental

Prácticas ESG y atracción de empleo

POLARIZING CORPORATIONS

TABLE 4. Job-Seekers' Preferences for Corporate ESG

	Interest (1)	Interest (2)	Interest (3)
ESG	0.098*** (0.026)	0.099*** (0.025)	0.085*** (0.020)
Ln(Wage)	1.117*** (0.031)	1.130*** (0.030)	1.205*** (0.026)
Nonwage Amenities	0.059*** (0.014)	0.060*** (0.014)	0.064*** (0.011)
Financial Strength	-0.003 (0.041)	-0.006 (0.040)	0.015 (0.032)
Observations	24,120	24,120	24,120
Individual FE	No	No	Yes
Strata FE	Yes	Yes	Yes
Controls			
Gender	No	Yes	-
Race	No	Yes	-
Age	No	Yes	-
Income	No	Yes	-
Employment Status	No	Yes	-
Political View	No	Yes	-

Notes: This table reports the regression coefficients for the following specifications. Column (1) specification: $Interest_{ij} = \alpha + \beta_1 ESG_{ij} + \beta_2 \ln(Wage_{ej}) + \beta_3 NW A_{ij} + \beta_4 FS_{ij} + e_{ij}$. Column (2) specification: $Interest_{ij} = \alpha + \beta_1 ESG_{ij} + \beta_2 \ln(Wage_{ej}) + \beta_3 NW A_{ij} + \beta_4 FS_{ij} + Demographic\ controls_i + e_{ij}$. Column (3) specification: $Interest_{ij} = \alpha + \beta_1 ESG_{ij} + \beta_2 \ln(Wage_{ej}) + \beta_3 NW A_{ij} + \beta_4 FS_{ij} + IndividualFE_i + e_{ij}$, i is the i -th individual and j is the j -th job posting rated by individual i . ESG is an indicator variable equal to one if the job posting displays at least one ESG sentence (see Appendix Table A21) or ESG certification (see Appendix Table A22). $\ln(Wage)$ is the natural logarithm of the monthly wage displayed in the job posting. $NW A$ is equal to the number of nonwage amenities. FS is an indicator variable equal to one if the job posting displays a signal of financial strength (see Appendix Table A19). Robust standard errors are reported in parentheses. * $p<0.1$; ** $p<0.05$; *** $p<0.01$.

Figure:

https://www.nber.org/system/files/working_papers/w31913/w31913.pdf

De la Teoría a la Evidencia

- ▶ Teoría ⇒ Proposiciones ⇒ Hipótesis
- ▶ Traducimos las proposiciones en hipótesis testeables

De la Teoría a la Evidencia

Representación de relaciones teóricas como un gráfico causal

Figure 4: Measurements and Hypotheses

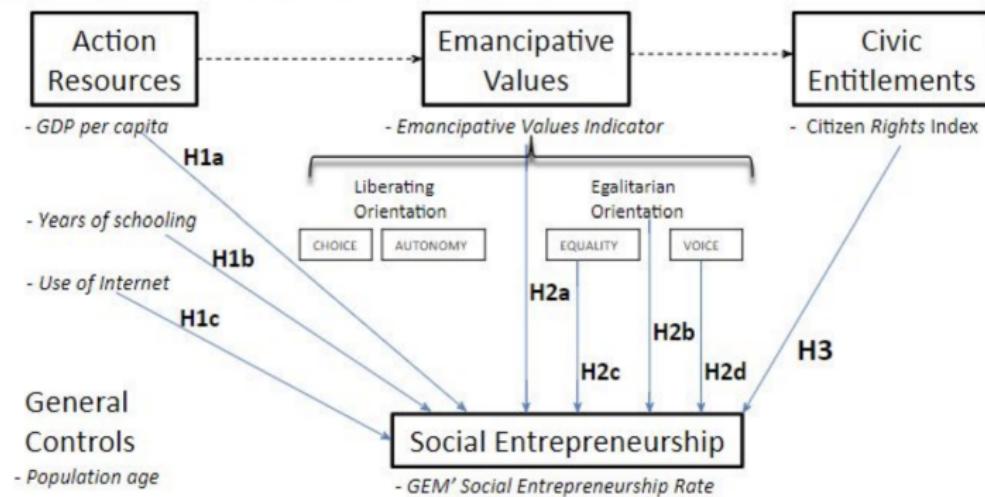


Figure: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3400084

De la Teoría a la Evidencia

Representación del Modelo como una Ecuación

$$SE_{i,t} = f(ActionResources_{i,t}, EmancipativeValues_{i,t}, CivicEntitlements_{i,t})$$

- ▶ Notar que i (un país) es la unidad de análisis
- ▶ SE_i es una medida del Emprendedorismo Social en el país i en el año t
- ▶ $f()$ es una función cuya forma desconocemos
- ▶ $Action Resources$ se aproximarán con el PBI per cápita, los años de escolaridad y el uso de internet.
- ▶ $Emancipative Values$ es una medida de los valores emancipativos en una sociedad (Welzel)
- ▶ $Civic Entitlements$ es aproximado por un indicador de derechos de la ciudadanía.

De la Teoría a la Evidencia

Representación como un Modelo de Regresión

$$SE_{i,t} = \beta_0 + \beta_1 ActionResources_{i,t} + \beta_2 EmancipativeValues_{i,t} + \beta_3 CivicEntitlements_{i,t} + \varepsilon_{i,t}$$

- ▶ El modelo de Regresión propone una forma lineal para la función en cuestión que podremos estimar.
- ▶ Notar la inclusión del término de error

Marketing: Elasticidad Precio y Estimación de Demanda

Objetivo: Estimar cómo el precio afecta la demanda de un producto.

Modelo:

$$\log(Q_i) = \beta_0 + \beta_1 \log(P_i) + \beta_2 X_i + \varepsilon_i$$

- ▶ Q_i : Cantidad vendida del producto i
- ▶ P_i : Precio del producto i
- ▶ X_i : Otras variables (promoción, publicidad, etc.)

Aplicación: Fijación de precios, planificación de promociones.

Referencia: Hanssens et al. (2001), *Market Response Models*

Finanzas: CAPM y Desempeño de Empresas

Objetivo: Comprender el riesgo sistemático y el retorno esperado.

Modelo CAPM:

$$R_{it} - R_{ft} = \alpha_i + \beta_i(R_{mt} - R_{ft}) + \varepsilon_{it}$$

- ▶ R_{it} : Retorno de la acción i en el tiempo t
- ▶ R_{ft} : Tasa libre de riesgo
- ▶ R_{mt} : Retorno del mercado

Aplicación: Evaluación de desempeño, decisiones de inversión.

Referencia: Fama y French (1993)

Comportamiento Organizacional: Brecha Salarial de Género

Objetivo: Analizar diferencias salariales según género.

Modelo:

$$\log(\text{Salario}_i) = \beta_0 + \beta_1 \text{Mujer}_i + \beta_2 \text{Educación}_i + \beta_3 \text{Experiencia}_i + \varepsilon_i$$

- ▶ **Mujer:** Indicador binario (1 si es mujer)
- ▶ Controles por capital humano

Aplicación: Análisis de discriminación, políticas de RRHH.

Referencia: Oaxaca (1973)

Operaciones: Tiempo de Espera y Satisfacción del Cliente

Objetivo: Medir cómo el tiempo de espera afecta la satisfacción.

Modelo:

$$\text{Satisfacción}_i = \beta_0 + \beta_1 \text{Espera}_i + \beta_2 \text{Resuelto}_i + \varepsilon_i$$

- ▶ Espera: Minutos de espera
- ▶ Resuelto: Indicador de si se resolvió el problema

Aplicación: Diseño de servicios, gestión de centros de atención.

Estrategia: Determinantes del Desempeño Empresarial

Objetivo: Entender qué factores explican el desempeño de una empresa.

Modelo:

$$\text{Desempeño}_i = \beta_0 + \beta_1 \text{I\&D}_i + \beta_2 \text{Participación}_i + \beta_3 \text{Industria}_i + \varepsilon_i$$

- ▶ Desempeño: ROA, ROI o utilidades
- ▶ Industria: Efectos fijos sectoriales

Aplicación: Estrategia competitiva, benchmarking.

Referencia: Rumelt (1991)

Notación para modelización usando regresión

$$\underbrace{Y_i}_{\begin{array}{l} \text{Variable} \\ \text{a explicar} \\ \text{o Dependiente} \end{array}} = \underbrace{\beta_0}_{\begin{array}{l} \text{Constante} \\ \text{cepto} \\ \text{o } \end{array}} + \beta_1 \cdot \underbrace{X_{1,i}}_{\begin{array}{l} \text{Variable} \\ \text{explicativa} \\ \text{(independiente)} \end{array}} + \beta_2 \cdot \underbrace{X_{2,i}}_{\begin{array}{l} \text{Variable} \\ \text{explicativa} \\ \text{(independiente)} \end{array}} + \cdots + \underbrace{\varepsilon_i}_{\text{Error}}$$

- ▶ $\beta_0, \beta_1, \dots, \beta_k$ son los coeficientes a ser estimados.
- ▶ Una vez estimados los denotamos con $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$.
- ▶ Los coeficientes medirán la contribución a la variable explicada por cada unidad de la variable explicativa.

Explicación y Predicción

▶ Explicación

- ▶ Los coeficientes estimados $\hat{\beta}_0, \hat{\beta}_2, \dots, \hat{\beta}_k$ son la base para la explicación.
- ▶ Si bien estrictamente hablando, los coeficientes solo capturan variaciones, y por lo tanto no representan necesariamente causalidad, son la base bajo las cuales buscaremos identificar efectos causales.

▶ Predicción

- ▶ La predicción se obtiene reemplazando los coeficientes estimados en la ecuación, y utilizando estos componentes para calcular el valor de que predice el modelo.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \cdots + \hat{\beta}_k X_{k,i}$$

Medidas de Bondad de Ajuste

R^2

- ▶ El R^2 es una medida de la bondad de ajuste del modelo en su conjunto.
- ▶ Se define como el cociente entre la varianza explicada por el modelo y la varianza total de la variable dependiente.
- ▶ Se interpreta como la proporción de la varianza de la variable dependiente que es explicada por el modelo.

Bondad de Ajuste: R^2

- ▶ Puesto que vale que

$$Y_i = \hat{Y}_i + \hat{\varepsilon}_i$$

- ▶ También vale que la varianza de Y_i es la suma de la varianza de la predicción y la varianza de los residuos.
- ▶ Esta igualdad nos permite definir una medida de la bondad de ajuste del modelo: el R^2 es la proporción de la varianza de la variable dependiente que es explicada por el modelo.

$$1 = \frac{Var(\hat{Y}_i)}{Var(Y_i)} + \frac{Var(\hat{\varepsilon}_i)}{Var(Y_i)}$$

$$R^2 = \frac{Var(\hat{Y}_i)}{Var(Y_i)} = 1 - \frac{Var(\hat{\varepsilon}_i)}{Var(Y_i)}$$

- ▶ Notar que es un valor entre 0 y 1.

Medidas de Bondad de Ajuste

Error Cuadrático Medio

- ▶ El error cuadrático medio (MSE) es otra medida de la bondad de ajuste del modelo en su conjunto.
- ▶ Es una medida que busca cuantificar la magnitud de los errores.

$$MSE = \frac{\sum \hat{\varepsilon}_i^2}{n - (k + 1)}$$

- ▶ El $-(k + 1)$ en el denominador es una corrección estadística al promedio, dada por el número de coeficientes estimados.
- ▶ La raíz cuadrada del MSE (o RMSE) es una medida que puede compararse en magnitud con la variable dependiente.